Predictive Analytics for Hospital Readmission Rates Using Patient Health Records

# Description

Analyze patient health records to identify key factors contributing to hospital readmissions within 30 days.

## Tech Stack:

* **SQL** (PostgreSQL / MySQL): for ETL, OLTP/OLAP schema design, data cleaning, and querying
* **Python**: for data wrangling (Pandas), visualization (Seaborn/Matplotlib), and modeling (Scikit-learn/XGBoost)
* **Jupyter Notebook**: for interactive analysis
* **Power BI / Tableau** (optional): for dashboarding
* **AWS RDS / EC2** (optional): to simulate a cloud-based deployment
* **Snowflake / MongoDB** (optional): for dimensional model vs NoSQL comparison

## Key Components:

**1. Data Modeling**

* Design an **OLTP schema** for raw patient records (diagnoses, admissions, procedures)
* Build a **Star Schema** for analytics (fact: readmissions, dimensions: patient, diagnosis, time)

**2. ETL/ELT Pipeline**

* Use Python scripts (with psycopg2 or SQLAlchemy) to:
  + Load CSV data into PostgreSQL
  + Clean and transform data (e.g., normalize ICD codes, filter outliers)
  + Create summary tables for analytics

**3. Exploratory Data Analysis (EDA)**

* Use Pandas and Seaborn to explore:
  + Readmission rates by diagnosis, age group, gender, etc.
  + Length of stay vs readmission correlation

**4. Predictive Modeling**

* Train ML models to predict likelihood of readmission:
  + Logistic Regression, Random Forest, or XGBoost
  + Evaluate with ROC-AUC, F1 Score, Precision-Recall

**5. Business Insights**

* Identify high-risk patient groups
* Recommend interventions (e.g., follow-ups, case management)

**6. Optional Advanced Layer**

* Use **Snowflake** to demonstrate cloud analytics benefits
* Use **MongoDB** for semi-structured data (e.g., doctor’s notes)

# Stage 1: Understanding the Problem

**Goal:** Predict whether a patient will be readmitted within 30 days based on their health records.

**Key Concepts:**

* **Readmission**: When a patient returns to the hospital soon after being discharged—often a quality-of-care issue.
* **Predictive Analytics**: Using historical data to predict future outcomes.

**Aims:**

* Identify risk factors for readmission.
* Help hospitals reduce preventable readmissions and save costs.

**Why It Matters:**

Hospitals are under pressure to reduce readmissions. High readmission rates often:

* Signal poor quality of care
* Lead to penalties from insurance programs (e.g., Medicare)
* Increase operational costs and patient dissatisfaction

So, a predictive system can:

* Flag high-risk patients for follow-up
* Support healthcare policymaking
* Improve resource allocation

**Key Questions We Want to Answer:**

1. **Who is at the highest risk of readmission?**
   * Elderly? Diabetic patients? Certain diseases?
2. **What patterns are common in readmitted patients?**
   * Long hospital stays? Specific procedures?
3. **Can we predict readmission using past data?**
   * Yes, through machine learning models.

# Stage 2: Data Modeling (SQL + DB Concepts)

**Concepts:**

* **OLTP (Online Transaction Processing)**: For storing raw, normalized data (e.g., patient, admission, diagnosis tables).
* **Star Schema / Dimensional Modeling (OLAP)**: For fast, analytical querying.

**Tasks:**

1. **Design OLTP tables**:
   * patients, admissions, diagnoses, procedures, medications
2. **Design a Star Schema**:
   * Fact Table: readmissions
   * Dimension Tables: dim\_patient, dim\_diagnosis, dim\_time

I’ll help you draw the ERD and write the SQL DDL to create these tables.

## Entity Relationship Diagram

A computer screen shot of a computer

AI-generated content may be incorrect.

## Entities & Their Key Attributes

**📋 1. patients**

* patient\_id (PK)
* first\_name
* last\_name
* gender
* date\_of\_birth

**📋 2. admissions**

* admission\_id (PK)
* patient\_id (FK)
* admission\_date
* discharge\_date
* length\_of\_stay
* readmitted (target variable: 0 or 1)

**📋 3. diagnoses**

* diagnosis\_id (PK)
* admission\_id (FK)
* diagnosis\_code (e.g., ICD-10)
* description

**📋 4. procedures**

* procedure\_id (PK)
* admission\_id (FK)
* procedure\_code
* description

**📋 5. medications**

* medication\_id (PK)
* admission\_id (FK)
* medication\_name
* dosage
* frequency

**6. doctors**

* doctor\_id (PK)
* first\_name
* last\_name
* specialty
* department

**🔗 7. admission\_doctors (Join Table)**

* admission\_id (FK)
* doctor\_id (FK)
* role (e.g., "Primary", "Consulting", etc.)

| **Parent Table** | **Child Table** | **Relationship** |
| --- | --- | --- |
| patients | admissions | 1-to-many |
| admissions | diagnoses | 1-to-many |
| admissions | procedures | 1-to-many |
| admissions | medications | 1-to-many |
| doctors | admission\_doctors | many-to-many (via join table) |
| admissions | admission\_doctors | many-to-many (via join table) |

# Stage 3: ETL (Extract, Transform, Load)

**Concepts:**

* **ETL**: Process of extracting raw data, cleaning and transforming it, then loading it into a data warehouse.
* **Python libraries**: pandas, sqlalchemy, psycopg2

**Tasks:**

* Use Python to:
  + Load data from CSV or an API
  + Clean nulls, standardize codes
  + Transform into analytics-ready format
  + Load into PostgreSQL

We’ll write ETL scripts step-by-step and run sample transformations.

## Extract

Extract data from source to pandas

**Tools:**

* Pandas
* SQLAlchemy

## Transform

* Involves cleaning up data for any missing or invalid values

# Stage 4: Exploratory Data Analysis (EDA)

**Concepts:**

* **Descriptive Statistics**: Mean, median, frequency, correlation
* **Visualizations**: Histograms, bar plots, heatmaps
* **Data Distributions**: Identifying skew, missing values

**Tasks:**

* Use pandas, matplotlib, and seaborn to:
  + Explore age vs readmission
  + Look at readmission by disease type
  + Analyze trends by length of stay

I'll show you how to build these charts and interpret them.

# Stage 5: Predictive Modeling (ML in Python)

**📘 Concepts:**

* **Supervised Learning**: Predict target variable (readmitted: yes/no)
* **Model Types**: Logistic Regression, Decision Trees, Random Forest, XGBoost
* **Evaluation Metrics**: Accuracy, Precision, Recall, AUC-ROC

**Tasks:**

* Split data into training/testing
* Build logistic regression baseline model
* Improve with tree-based models
* Use confusion matrix and ROC curve to evaluate

We’ll go hands-on with Scikit-learn and XGBoost.

# Stage 6: Business Insights & Reporting

**📘 Concepts:**

* **Actionable Insights**: Translate results into business strategy
* **Dashboards**: Optional Power BI or Tableau
* **Data Storytelling**: Presenting results in a clear, persuasive way

**Tasks:**

* Highlight most at-risk patient groups
* Suggest policy actions (e.g., post-discharge follow-up)
* Create a final report or dashboard

# Outcome:

* Full data pipeline from raw data to actionable prediction
* You’ll show SQL, Python, data modeling, EDA, ML, and reporting
* A portfolio-ready project to impress professors or employers